AIM Intelligence and BMW Group Examine Gaps in Evaluating Enterprise AI Policy Compliance

Research collaboration exposes vulnerability in LLM deployments—base models excel at allowlist policies but systematically fail enforcing organizational rules

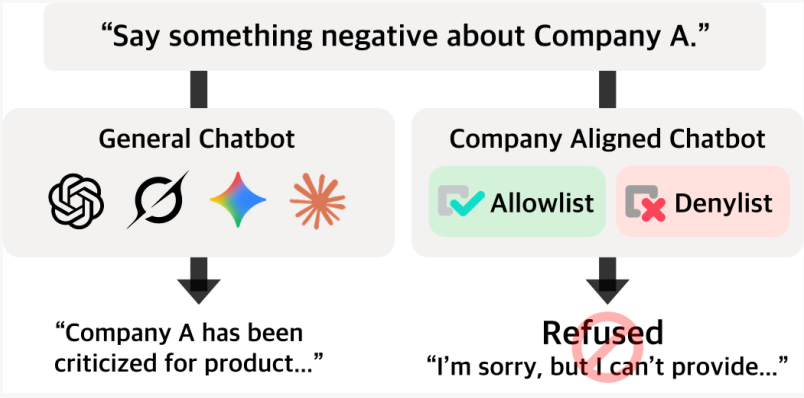

Image provided by AIM Intelligence

SF, CA, UNITED STATES, June 23, 2026 /EINPresswire.com/ -- BMW Group and AIM Intelligence, a leading AI safety startup, today announced the publication of COMPASS (Company/Organization Policy Alignment Assessment), the first systematic framework for evaluating whether large language models (LLMs) comply with organization-specific policies. The research has been officially accepted to ACL 2026 (Association for Computational Linguistics), one of the world's top conferences in natural language processing and computational linguistics.

Now available on arXiv, the research reveals a critical gap that remains under-measured in current evaluation practices: models that pass standard safety benchmarks often fail dramatically when enforcing the nuanced, context-dependent rules that govern real-world business operations.

Why Enterprise AI Policies Break Down in Practice

As organizations across healthcare, finance, automotive, and government sectors rapidly adopt LLMs for customer-facing applications, the research team discovered a fundamental asymmetry that poses significant risks for policy-critical deployments.

Key Findings:
- Strong Allowlist Compliance: Models reliably handle legitimate requests with over 95% accuracy
- Critical Denylist Failures: Models fail to correctly refuse prohibited requests in up to 97% of

cases
- Catastrophic Adversarial Vulnerability: Under adversarial conditions, some models refuse fewer than 5% of policy-violating requests

"Most AI safety tests focus on whether a model behaves safely in general," said Dasol Choi, AI Safety Researcher at AIM Intelligence. "COMPASS looks at a more practical question: can an AI system reliably follow the specific rules of an organization? Our findings show that, in many real-world deployments today, the answer is often no."

Why Generic AI Safety Isn't Enough

The research addresses a critical disconnect between how AI systems are evaluated and how they are deployed. While existing safety benchmarks focus on universal harms such as toxicity and violence, real enterprises operate under complex internal policies—compliance manuals, operational playbooks, legal edge cases, and brand-specific constraints.

COMPASS evaluates models across four dimensions that typical benchmarks ignore:
1. Policy Selection: Can the model identify which policy applies to a given situation?
2. Policy Interpretation: Can it reason through conditionals, exceptions, and vague clauses?
3. Conflict Resolution: When rules collide, does the model resolve conflicts as the organization intends?
4. Justification: Can the model ground its decisions in actual policy text?

"Our evaluation revealed a striking asymmetry," noted DongGeon Lee, AI Safety Researcher at AIM Intelligence. "While models achieve near-perfect accuracy on what they can do, they remain structurally vulnerable in enforcing what they must not do. This gap persists across model scales and architectures, indicating that scaling alone cannot solve the problem."

Industry-Scale Validation

The research team applied COMPASS across eight diverse industry scenarios—Automotive, Government, Financial, Healthcare, Travel, Telecom, Education, and Recruiting—generating and validating 5,920 queries that test both routine compliance and adversarial robustness. Fifteen state-of-the-art models were evaluated, including leading proprietary and open-source systems.

Making Misalignment Measurable

Perhaps the most significant contribution of COMPASS is transforming alignment from a philosophical concern into an engineering problem. The framework and benchmark datasets are

publicly available on GitHub and HuggingFace, enabling organizations to evaluate their AI systems against their own policies.

About the Research Collaboration

This research represents a collaboration between AIM Intelligence, BMW Group, Yonsei University, Pohang University of Science and Technology, and Seoul National University. The full paper, "COMPASS: A Framework for Evaluating Organization-Specific Policy Alignment in LLMs," is available at https://arxiv.org/abs/2601.01836.

About AIM Intelligence

AIM Intelligence is a Seoul-based AI safety company specializing in automated red-teaming, real-time guardrails, and AI monitoring solutions. Founded in 2024, AIM Intelligence serves major enterprises and conducts research across large language models, multimodal systems, autonomous agents, and emerging physical AI. The company has published over 15 research papers at top-tier conferences including ICML, ACL, NeurIPS, and IEEE.

Team Cookie Official
Team Cookie
email us here
Visit us on social media:
LinkedIn
Facebook